RS5M and GeoRSCLIP: A Large-Scale Vision- Language Dataset and a Large Vision-Language Model for Remote Sensing

Year: 2023 | *Citations: 130* | *Authors: Zilun Zhang, Tiancheng Zhao, Yulong Guo*

Abstract

Pretrained vision-language models (VLMs) utilizing extensive image–text paired data have demonstrated unprecedented image–text association capabilities, achieving remarkable results across various downstream tasks. A critical challenge is how to make use of existing large-scale pretrained VLMs, which are trained on common objects, to perform the domain-specific transfer for accomplishing domain-related downstream tasks. In this article, we present an image–text paired dataset in the field of remote sensing (RS), RS5M, which has 5 million RS images with English descriptions. The dataset is obtained from filtering publicly available image–text paired datasets and captioning label-only RS datasets with pretrained VLM. These constitute the first large-scale RS image–text paired dataset. Additionally, we present GeoRSCLIP by fine-tuning (FT) or applying parameter-efficient FT (PEFT) methods to the CLIP model using RS5M. Experimental results show that our proposed dataset is highly effective for various tasks, and our model GeoRSCLIP improves upon the baseline or previous state-of-the-art model by 3%–20% in zero-shot classification (ZSC) tasks, 3%–6% in RS cross-modal text–image retrieval (RSCTIR) and 4%–5% in semantic localization (SeLo) tasks. Dataset and models have been released in: https://github.com/om-ai-lab/RS5M.